Comp 3202 Report

Question 1

As seen in the graph on the right. While testing various values of k on the euclidean distance model KNN, increasing values of k increase the capacity of the model, and lower training and testing error to a point before the error increases again.

Region 1:
For high values of k, the model shows obvious underfitting, with a capacity too low to accurately classify any data. The bias of the model is high in this region, and variance is low.

Region 2
As the value of k decreases, the model begins to more accurately fit to the data. Higher capacity values allow the model to correctly classify more points, dropping both testing and training error. In this region variance is steadily increasing, and bias decreasing as the model becomes increasingly specialized to the training data.

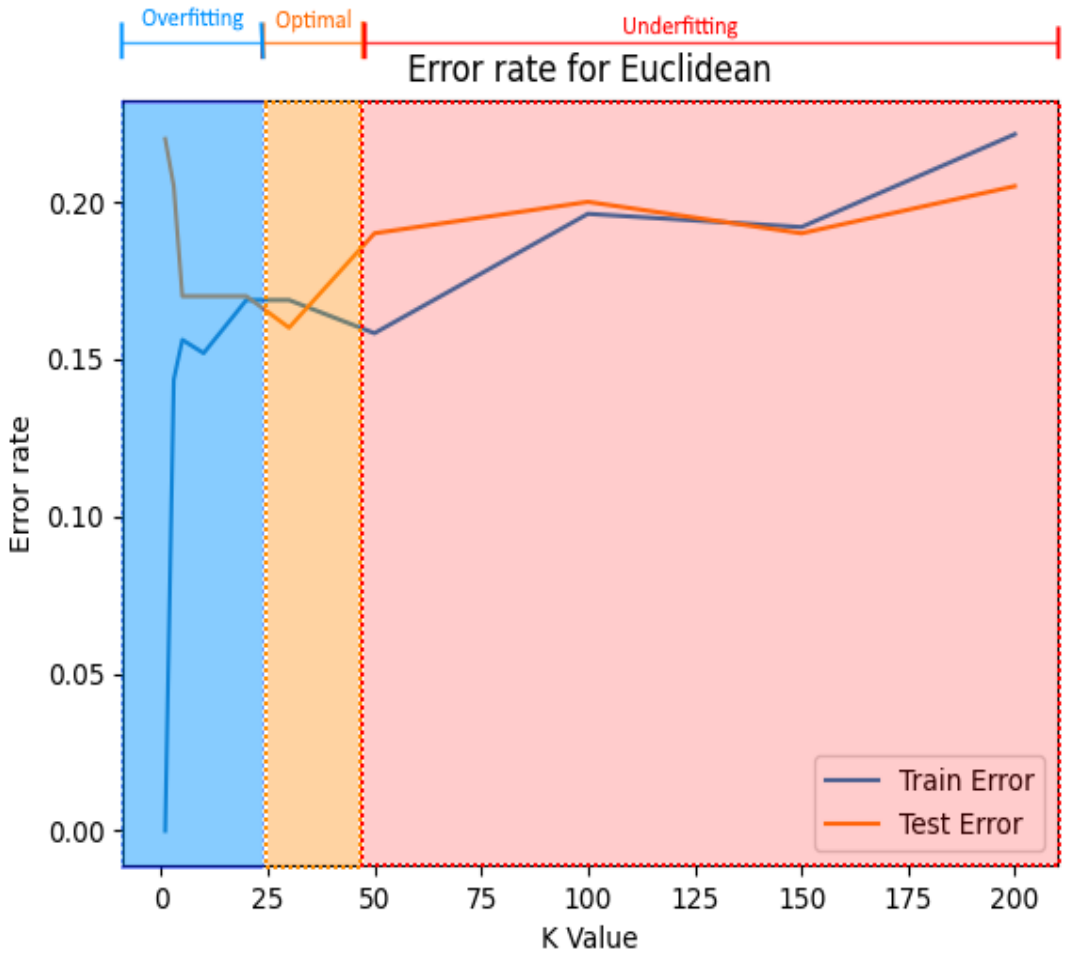
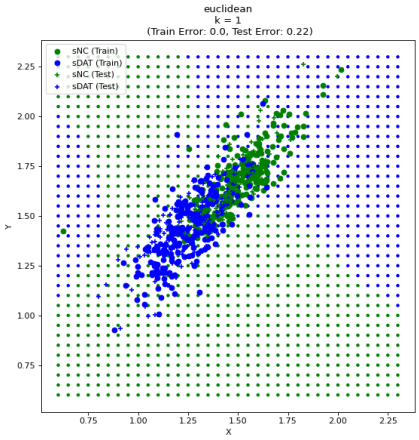
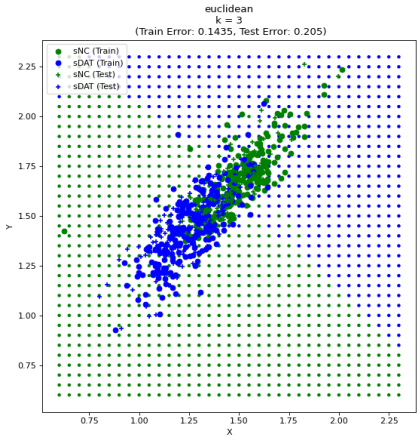


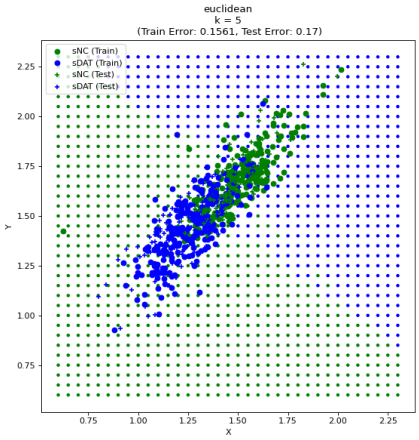
Figure 1: a graph of testing error vs the hyperparameter K. Regions of overfitting, underfitting, and optimal fitting have been roughly highlighted.

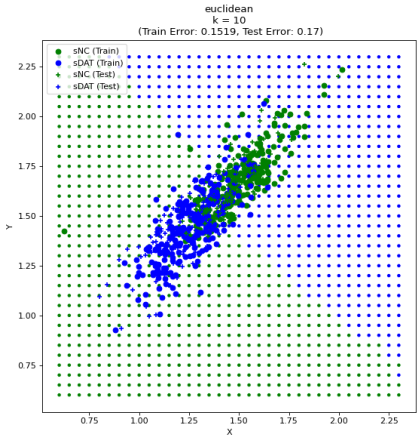
Region 3:
With sufficiently low values of k, the model passes the threshold into overfitting. At this point the gap between training and test error increases rapidly, as the model has overspecialized to the training data. There is almost no bias in this region and very high variance. The model only accurately classifies points from the original training data set.

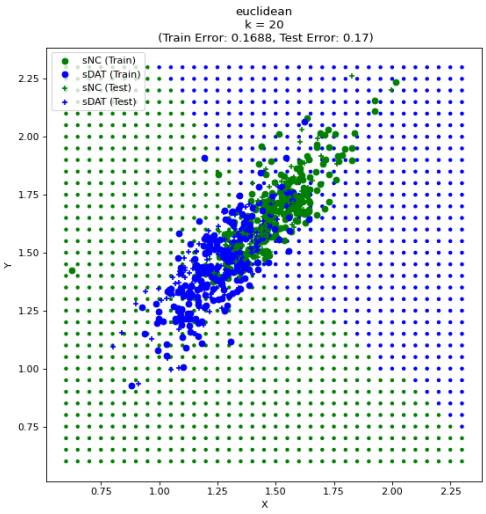
Euclidean K Value Plots

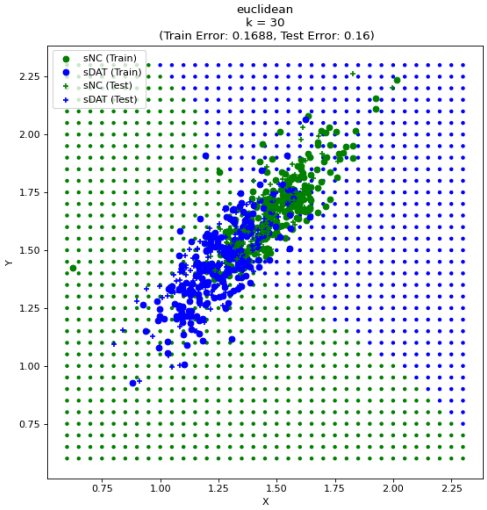


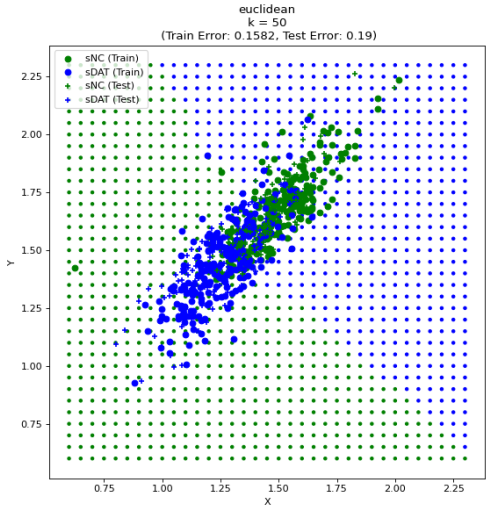


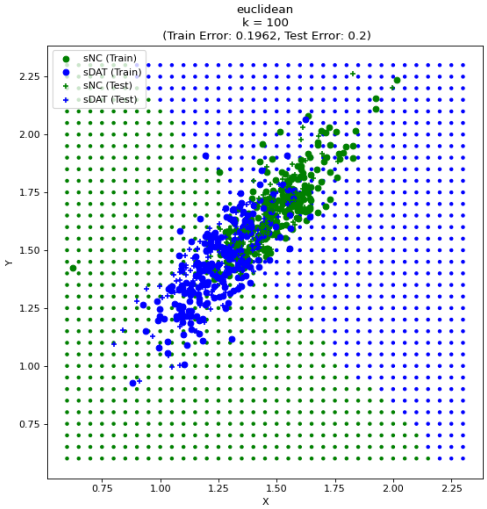


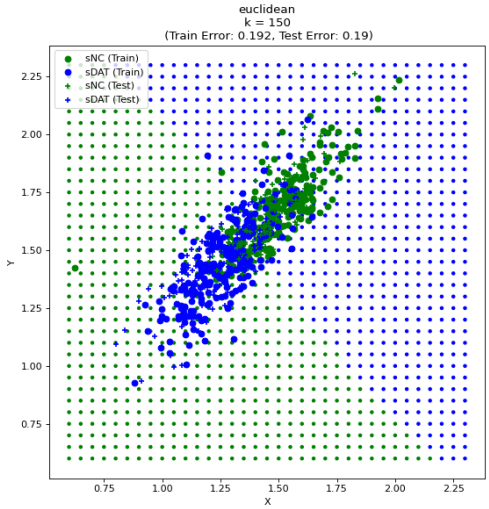


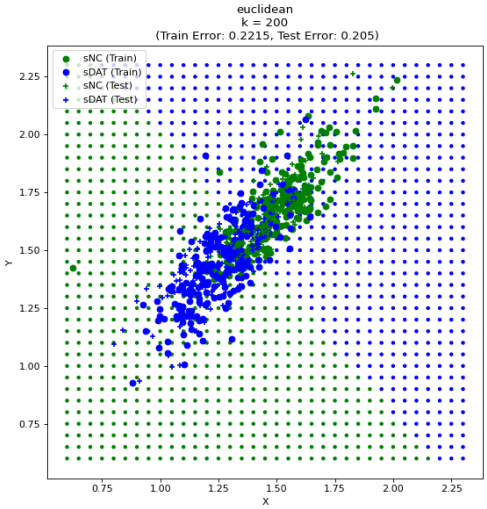



euclidean
k = 30
(Train Error: 0.1688, Test Error: 0.16)
sNC (Train)
sDAT (Train)
sNC (Test)
sDAT (Test)
X
Y


euclidean
k = 50
(Train Error: 0.1582, Test Error: 0.19)
sNC (Train)
sDAT (Train)
sNC (Test)
sDAT (Test)
X
Y

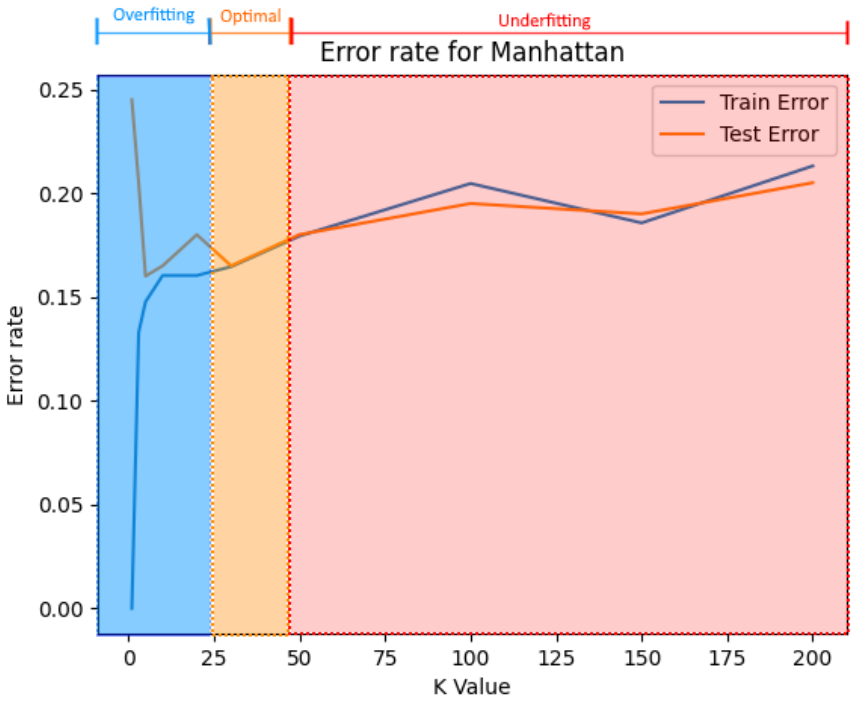

euclidean
k = 100
(Train Error: 0.1962, Test Error: 0.2)
sNC (Train)
sDAT (Train)
sNC (Test)
sDAT (Test)
X
Y


euclidean
k = 150
(Train Error: 0.192, Test Error: 0.19)
sNC (Train)
sDAT (Train)
sNC (Test)
sDAT (Test)
X
Y


euclidean
k = 200
(Train Error: 0.2215, Test Error: 0.205)
sNC (Train)
sDAT (Train)
sNC (Test)
sDAT (Test)
X
Y

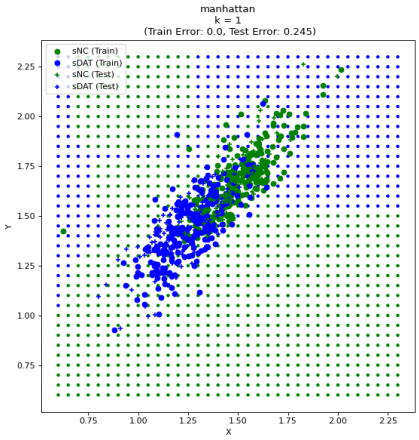
Question 2

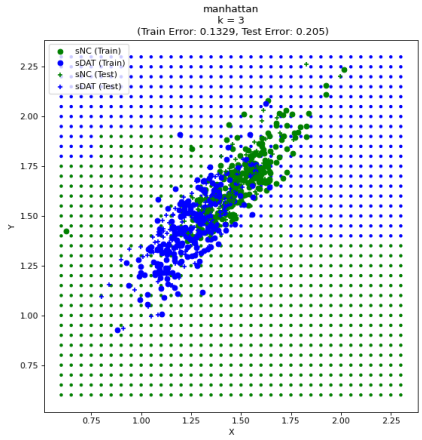
The Manhattan classifier produces very similar results to the Euclidean classifier in terms of shape. Like the Euclidean classifier, as the value of k decreases, the model moves from underfitting, to overfitting, with the capacity, and variance increasing and the bias decreasing.

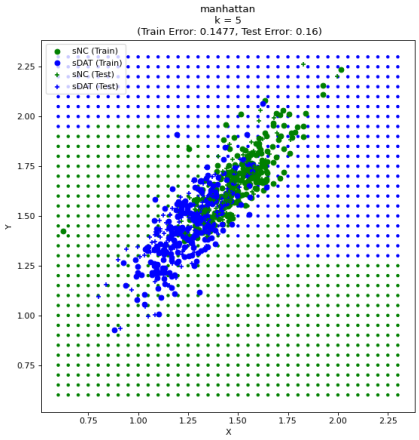


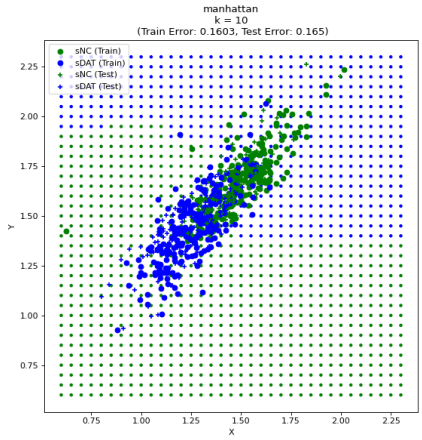
Notably, the main difference between the Manhattan and Euclidean classifiers was that the Manhattan achieved superior test error rates overall. Though both models tied for the lowest overall error rate, with Euclidean having its lowest error rate for k = 30, and Manhattan for k = 5. From this data we can conclude that the Manhattan distance metric will likely lead to lower overall test error.

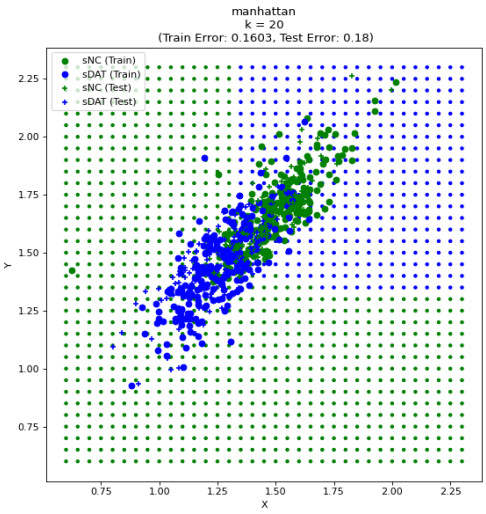
Manhattan k value plots

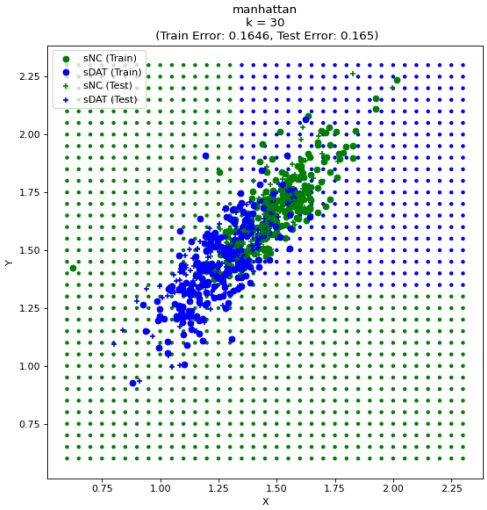


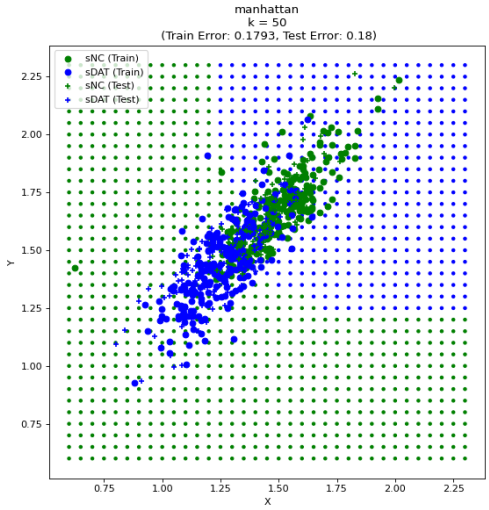


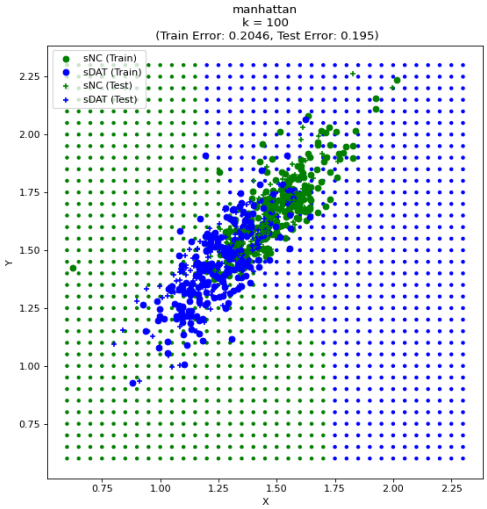


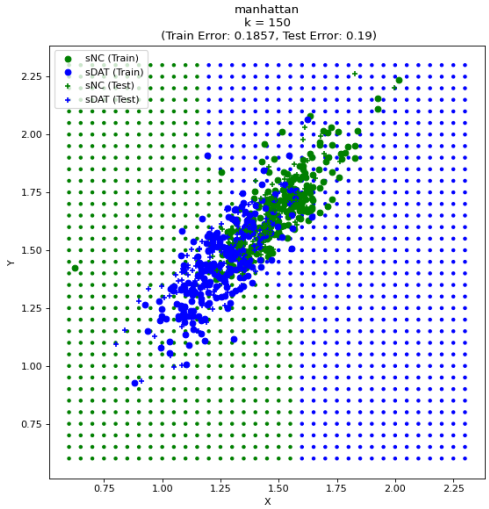


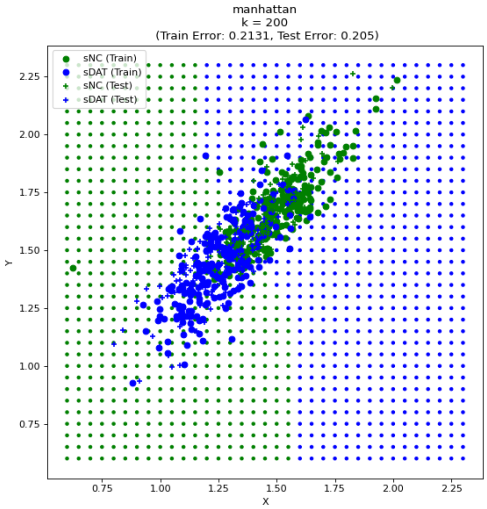



manhattan
k = 30
(Train Error: 0.1646, Test Error: 0.165)
sNC (Train)
sDAT (Train)
sNC (Test)
sDAT (Test)
X
Y

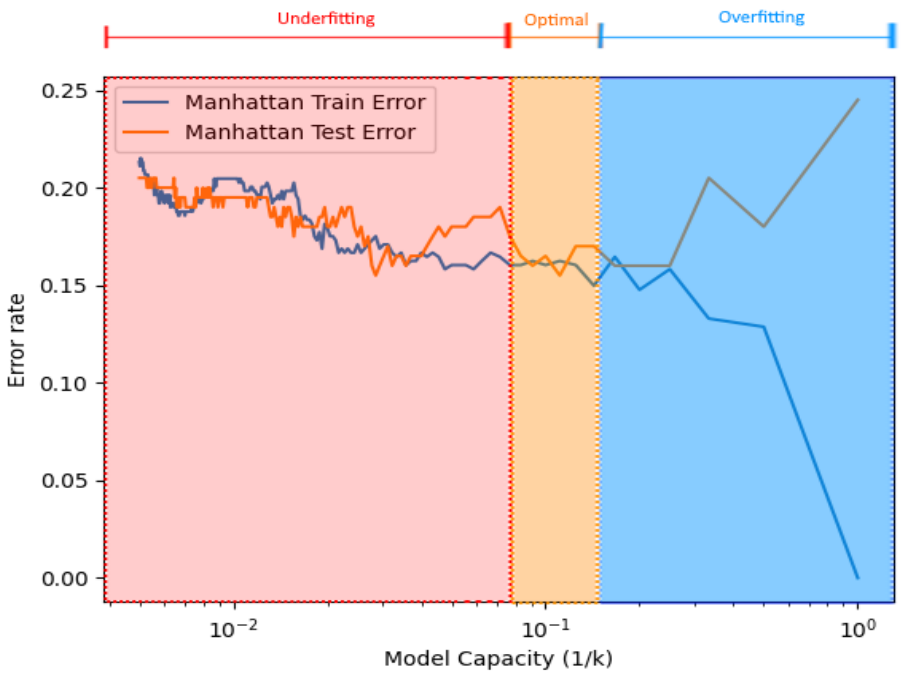

manhattan
k = 50
(Train Error: 0.1793, Test Error: 0.18)
sNC (Train)
sDAT (Train)
sNC (Test)
sDAT (Test)
X
Y


manhattan
k = 100
(Train Error: 0.2046, Test Error: 0.195)
sNC (Train)
sDAT (Train)
sNC (Test)
sDAT (Test)
X
Y


manhattan
k = 150
(Train Error: 0.1857, Test Error: 0.19)
sNC (Train)
sDAT (Train)
sNC (Test)
sDAT (Test)
X
Y


manhattan
k = 200
(Train Error: 0.2131, Test Error: 0.205)
sNC (Train)
sDAT (Train)
sNC (Test)
sDAT (Test)
X
Y

Question 3



The above graph shows the test error rate versus the model capacity. It can be visibly separated into 3 regions.
Region 1:
With a very low capacity, the model has very high bias and low variance. It has relatively high testing and training error rates, and the error rates are roughly the same for both groups. This is because the model is underfitted, and cannot accurately predict data points.

Region 2:
At sufficiently high capacity, the model reaches a good balance of variance and bias. In this region, which is roughly marked as the orange region on the above graph, the model can still generalize to new data points while still properly fitting itself to data. Note that this is the lowest point in the test error rates, and that test and training error rates coincide with little separation.

Region 3:
Increasing model capacity further into region three sees a sharp decline in test accuracy, that trends towards 0. However there is, in this region, an increasing gap between the test and training error rates. As the bias declines sharply and model variance increases, the model is now overfitted to the training data, and can now only accurately classify points that are originally in its training data.